A STUDY OF TECHNIQUES FOR DATA MINING AND FRAMEWORK FOR AGRICULTURE PREDICTION SYSTEM

UNIVERSITY LOGO

A
Synopsis
Submitted to the
Jayoti Vidyapeeth Women's University
for the Degree
Of
Doctor of Philosophy(Computer Science)

Supervised by

DR. SHOBA LAL

Submitted by

Ankit Meena
(JVR-II/22/11002)

FACULTY OF EDUCATION & METHODOLOGY
DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY
JAYOTI VIDYAPEETH WOMEN'S UNIVERSITY
JAIPUR (RAJASTHAN), INDIA

AUGUST 2023

Introduction

The world of now is filled with data. People who live in the digital age need instant access to reliable information. Data is widespread. A huge amount of data is generated every day. Data are being gathered and built up at an unexpected rate as a consequence of the constantly growing volumes of digital data. Finding useful information from these data is a fundamental question that is addressed to everyone. In order to extract useful information from the huge quantity of data that is currently available, data mining, a relatively new area of research, is essential. Whenever simple research and reporting techniques cannot be used, data mining can be used.

1.1 Data Mining

Every day, huge volumes of data are produced. Due to the ever-increasing volumes of digitaldata, data are being recorded and collected at an unexpected rate. An important issue that is addressed to everyone is how to derive information that is meaningful from this facts. Data mining, a relatively recent field of research, is crucial in order to extract valuable informationfrom the enormous amount of data now available.

Data mining is an emerging technique for processing large amounts of raw data to extract relevant information. It helps in classifying data, categorizing data, forecasting or estimating future values, determining correlations, and seeing patterns in the dataset. The unexceptionallylarge amounts of data that are present in daily life have made data mining a vital component.

A step in the bigger process of knowledge discovery from data is called data mining (data). Data mining is one step in the KDD process, which aims to extract useful and valuable knowledge from data. Preprocessing is used in data mining on the first gathered enormous datasets related to any topic or industry. Preprocessing involves the process of modifying datato make it suitable for data mining techniques to be used on it. Preprocessing comprises cleansing, transforming, and summarizing data, among other things.

Data is changed into the format required for the analysis. The population as a whole is represented by the dataset used. As a result, proper sampling is crucial to obtaining reliable conclusions from the dataset. Data warehouses are the biggest types of data storage. Any field's previous information can be retrieved in the Datawarehouse.

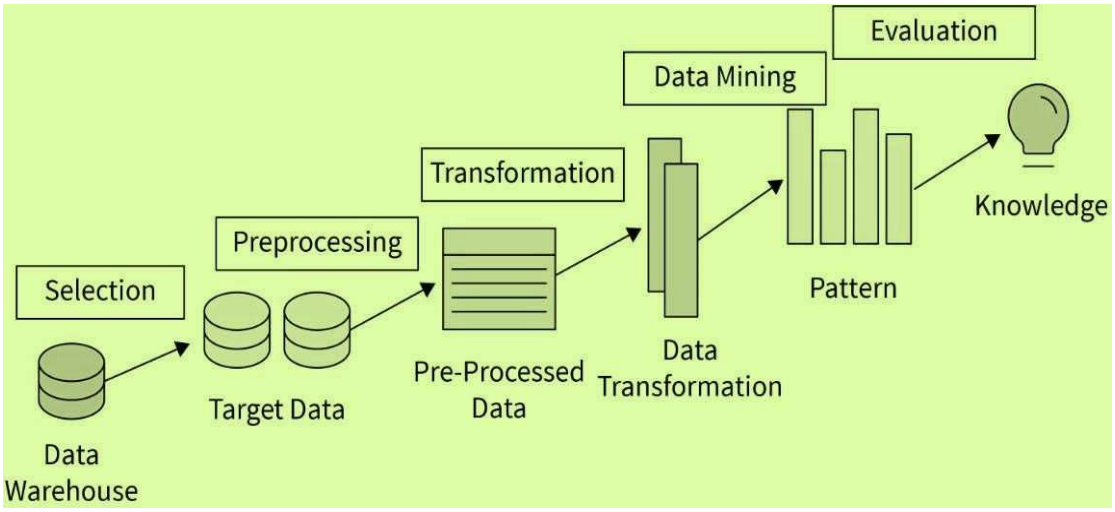

Figure 1.1: KDD Process

A KDD process is illustrated in figure 1.1. The steps in KDD process includes the collection ofdata, preprocessing of those data, transforming the data to the needed form, applying Data mining technique(s) on the transformed data and pattern analysis.

1. **Data Collection**: Data can be collected from the real world or from the various onlinerepositories or online sources depending on the nature of application.
2. **Data Preprocessing**: The data collected is in raw format which needs to be convertedto a perfect data. The data should be cleaned by removing noisy data, filling blank data, etc., so that finding the data patterns for the prediction analysis will be accurate.
3. **Data Transformation**: The pre-processed data is transformed and normalized so thatdata analysis can be done.

4. **Data mining**: Appropriate Data mining technique can be applied on the normalizeddata to get the future perspective patterns from the data collected.
5. **Evaluation**: The model generated by the Data mining technique is evaluated based oncertain metrics.

1.2 Data mining in Agriculture

The technique of extracting meaningful patterns from massive amounts of data is known as data mining. A lot of useful statistics are provided by the agricultural sector. Information technologyin agriculture can aid the farmer in decision-making and increase production more effectively. Farmers not only grow crops, but the cultivation also yields a significant amount of useful information.

There is a huge amount of agricultural data. Data mining thus offers a means of converting this huge volume of data into informative information. Basic information about their soils and the crops that can be cultivated in various types of soil is insufficient among farmers. They are unable to select the best crop for cultivation, forecast the yield, determine the fertilizers to use,forecast the market price, or identify agricultural illnesses. This could cause them to suffer a great loss. As a result, farmers believe that agriculture is an unreliable source of income or employment. The Data Mining model for recommending **profitable crops based on the soil for cultivation** by forecasting the **crop yield, crop price, fertilizers to be used, and identifying diseases,** will be used to solve all of these problems.

1.3 Role of Data mining in the prediction of crop, crop yield, fertilizer, crop price and crop disease

Extracting hidden predictive and useful information from large databases is called Data mining. It is an effective tool that has the ability to help a lot of people in deciding on what is most important in their data warehouses. The objective of the data mining process is to extract knowledge from an existing dataset and change it into a human readable a form for future use.

It is the technique of examining data from various perspectives and converting it into informative analysis. Any form of data can be analyzed via data mining without limitations. By detecting future patterns and behaviors, data mining enables firms to make proactive, knowledge-driven decisions.One of the areas where Data mining could provide potential

support is Agriculture. Vast amount of agricultural data is available. Similar to other sectors, the amount of agriculture data is increasing daily. Implementing information technology in agriculture can alter the decision-making process and help farmers produce better-quality food. Data mining is essential for decision-making in the agricultural sector on various kinds of challenges.

2. Objectives of the Research Work

The main objective of this research work is to create a framework to assist farmers in

- **Selecting** the crop for cultivation based on

i. Soil nutrients
ii. Crop price
iii. Crop yield
iv. Fertilizer

- **Identifying** the cultivated crop disease thereby preventing crop loss

To achieve the above objectives, the following sub goals are to be achieved:

- To assist farmers in identifying the crops for cultivation based on the soil nutrients availability, predicted crop yield, predicted crop price and fertilizers to be applied.
- To analyze the performance of different Classification algorithms for crops prediction based on various metrics on the

- Original soil dataset
- Dimensionality reduced soil dataset.
- Soil datasets with different sizes

- To assist farmers in predicting the yield of the crop to be cultivated based on previous yield history, the current season and climate.
- To analyze the existing algorithms for yield prediction and propose a new algorithm.
- To help farmers increase crop yield by applying the right amount of fertilizer that corresponds with the soil nutrient and crop being cultivated.

- To analyse the performance of different Clustering algorithms based on various metrics and propose a new algorithm for fertilizer suggestion.

3. RESEARCH QUESTIONS/HYPOTHESIS

- How can predicted crop yield help farmers in selecting the appropriate crops?
- How does the combination of soil nutrients availability, predicted crop yield, predictedcrop price, and fertilizers contribute to informed crop selection decisions?
- How can farmers use the information about previous yield history to predict crop yield?

4. RESEARCH METHODOLOGY

The main purpose of this work is to help farmers by providing necessary information at the right time so that yielding a better quality and quantity of the production, thereby improving their livelihood.

4.1 Steps will take in this research

In this research work, data will be collected from the physical records of government (soil centres) and from the internet. These data will be pre-processed and converted to .CSV format.Five main problems in agriculture are dealt here. Based on the nature of the problem, classification or clustering or time series or regression model will be designed. Some new better algorithms are proposed and new techniques for better performance are proposed. Finally theseproblems (modules) are put together under a framework to give a better solution to the farmers.

4.2 Proposed model

In the proposed model, the user (generally farmer) can sit at his/her place and request for information by giving the needed data to the server using Internet. First, from the soil data of the user's farm, in the Crop Prediction module, the set of crops to be cultivated is to be predicted using a better classifier algorithm. For those crops, based on the current season and weather condition prevailing in the farm, their yield is to be estimated using the to-be proposed Modified Gradient Boosting Regression algorithm in the Crop Yield Prediction module based on the pricehistory of the crop, the crop price is to be estimated using the to be proposed

Enhanced Exponential Smoothing algorithm in the Crop Price Prediction module and based on the soil nutrients availability, appropriate fertilizer for the corresponding nutrients and crop data, fertilizers to be applied is to be suggested using the to-be proposed Modified Agglomerative Hierarchical Clustering algorithm in the Fertilizer Suggestion module. These estimated yield, price and predicted fertilizers to be applied for the suggested crops are to be given as response tothe user. Based on the response, the most profitable crop for cultivation can be chosen by the user.To identify the disease that has attacked the crop cultivated, the user can request by giving the various characteristics of the crop grown. Using the input in the Crop Disease Prediction module,the affected disease is to be predicted using the best Oversampling technique on Linear SVM.

4.3 Experimental design of the study

In this work, all the problems are to be implemented using PYTHON with the dataset in .CSV format. Based on the nature of the problem and dataset, classification or clustering or time series or regression model is designed. The performance of the algorithms are analysed based on the corresponding metrics. Some new algorithms and techniques those are better than the existing one are proposed.

The steps involved are

- Collection of data
- Preprocessing the data
- Identifying the model to be designed
- Applying the existing algorithms
- Proposing a new better algorithm or technique
- Finally giving a framework

4.4 Datasets will be used in this study.

A dataset is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of arrays to a database table. In a tabular dataset, each column describes a particular variable, and each row corresponds to the fields of the dataset. The most supported file type for a tabular dataset is "Comma Separated File," or CSV. Some datasets are available online. In this work some online and offline datasets will be used.

5. Literature survey

Payal et al. [1] analyzed and compared the behaviour of different kinds of Classification algorithms on medical dataset taken from literature. On three different datasets, Naive Bayesian,Decision Tree Induction, Multilayer Perceptron and logistic regression were applied. The Classification techniques were compared based on the metrics TPrate, FPrate, Accuracy Precision, Accuracy Recall, Classification accuracy and the Execution time in seconds. They have concluded that MLP is better than others and more accurate but it takes more time.

E. Manjula and S. Djodiltachoumy [2] had analyzed soil nutrients by classifying the soil using Naive Bayes, Decision Tree and Hybrid approach of Naive Bayes and Decision Tree. The performances of these algorithms were compared based on the two factors, accuracy and Execution time, and concluded that the hybrid Classification algorithm outperformed the other two methods. In the report [3], the performance of Decision Tree, Naïve Bayes and kNN algorithms for predicting breast cancer with the help of WEKA was compared, based on the parameters lowest computing time and accuracy; and concluded that Naive Bayes is a better one.

Meraj et al. [4] had evaluated the performance of Naive Bayes, Logistic Regression,Decision Tree and Random Forest for predicting the diabetic patients based on Correctly Classified Instances, Incorrectly Classified Instances, Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Mean Square Error (RMSE), Root Relative Squared Error (RRSE) and Confusion Matrix. It was concluded, that Logistic Regression has more Correctly Classified Instances with low RMSE and RRSE whereas Naive Bayes has lowRAE and MAE.

A web-based solution for the soil testing laboratories was framed by Hari andJayasudha [5]. It provides information like soil testing code, fertilizer and next crop that can be cultivated, and if the farmer specifically needs a crop to be cultivated, information about the fertilizer needed for that crop is provided. This paper has evaluated and compared three Classification techniques (Naive Bayes, J48 (C4.5) and JRip) in data mining on the basis of Execution time, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Tenfold cross-validation was used in theexperiment. It was concluded that J48 (C4.5) model was found to be the best classifier for soil samples.

The performance of Naive Bayes, Bayesian Network, Naive Bayes updatable, J48 and Random Forest on soil dataset based on Relative Absolute Error, Mean Absolute Error, Correctly Classified Instances and Cohen’s Kappa Score Statistics was analyzed and compared by

Ramesh and Ramar [6]. They have concluded that among the taken Classificationtechniques Naive Bayes classifier is the efficient Classification technique.

Jay et al. [7] applied Naive Bayes, J48 and JRip on soil samples. The metric used for analysis were Correctly Classified Instances, Incorrectly Classified Instances, Accuracy and Mean Absolute Error. J48 was found to give the best result.

Kiran et al. [8] proposed the use of Data mining techniques to provide farmers with recommendations for crops, crop rotation and identification of appropriate fertilizer. For the recommendation of crops, Random Forest algorithm with rating system is proposed. Place and climate are given as inputs to the Random Forest algorithm.

A price prediction model based on Decision Tree techniques, named as Hybrid Association rule-based Decision Tree (HADT) algorithm was proposed by S. Rajeswari and K.Suthendran [9]. They have used agriculture daily crop price dataset of Virudhunagar District, Tamilnadu with the attributes commodity code, year, month, commodity names, commodity price, price prediction and class for the year 2016 - 2018. The performance of the algorithm was compared with the performance of other Decision Tree techniques based on the metrics, accuracy, rate of error and time of execution; and concluded that the HADT forecast model is better than other Decision Tree models and precise to predict agricultural product prices.

An Artificial Neural Network (ANN) works in processing information similar to the biological nervous systems, such as the brain. An ANN through a learning process is configured for a specific application, such as data classification or pattern recognition. From complicated or imprecise data, ANN can derive meaning. Neural networks can be used to extract patterns and detect trends that either humans or other computer techniques find it to be complex. Once trained, the neural network is the "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions [10].

In the survey paper, Karandeep [11] proposed to apply machine learning techniques for agriculture. It recommended ANN, kNN and Decision Trees based on various factors like type of soil and its composition, climate, geography of the region, crop yield, market prices etc.

REFERENCES

1. Payal Pahwa, Manju Papreja and Renu Miglani, "Performance Analysis of Classification Algorithms", *International Journal of Computer Science and Mobile Computing*,Vol. 3(4), pp 50-58, 2014.
2. E. Manjula and S. Djodiltachoumy, "Data Mining Technique to Analyze Soil Nutrients Based on Hybrid Classification", *International Journal of Advanced Research in Computer Science*, Vol. 8(8), pp 505-510, 2017
3. Chintan Shah and Anjali G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction", *International Conference on Computing, Communications and Networking Technologies*, pp 1-4, 2013.
4. Meraj Nabi, Abdul Wahid, and Pradeep Kumar, "Performance Analysis of Classification Algorithms in Predicting Diabetes", *International Journal of Advance Research in Computer Science,* Vol. 8(3), pp 456-461, 2017.
5. S.Hari Ganesh, and Jayasudha, "Data Mining Technique to Predict the Accuracy of the Soil Fertility", *International Journal of Computer Science and Mobile Computing*,Vol. 4(7), pp 330-333, 2015.
6. V.Ramesh and K.Ramar, "Classification of agricultural land soils: A data mining approach", *Agricultural Journal,* Vol. 6(3), pp 82-86, 2011.
7. Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade and Vahida Attar, "Soildata analysis using classification techniques and soil attribute prediction", *International Journal of Computer Science Issues*, Vol. 9(3), pp 1-4, 2012
8. Kiran Shinde, Jerrin Andrei and Amey Oke, "Web Based Recommendation System for Farmers", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 3(3), pp 1444-1448, 2015
9. S.Rajeswari and K.Suthendran, "Developing an Agricultural Product Price Prediction Model using HADT Algorithm", *International Journal of Engineering and Advanced Technology*, Vol. 9(1S4), pp 569-575, 2019.
10. N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6(3), pp 59-65, 2016.
11. Karandeep Kaur, "Machine Learning: Applications in Indian Agriculture", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5(4), pp 342-344, 2016.